Tools for Polyploids Workshop
Computational Support

SNP and Dosage Calling

Cristiane Taniguti
Gabriel Gesteira
Jeekin Lau
Maria Caraza-Harter

TOOLS FOR POLYPLOIDS

This project is funded by USDA NIFA Specialty Crop Research Initiative Award # 2020-51181-32156 (09/01/20 - 08/31/24)

Getting Prepared for the Workshop

- Polyploids
- Molecular Markers
- Genome variations – applications
  - Quantitative traits mapping
  - Genome Wide Association studies
  - Phenotypic predictions - Genome Selection
  - Evolution and diversity studies
  - Gene expression studies

TOOLS FOR POLYPLOIDS

Genome variations

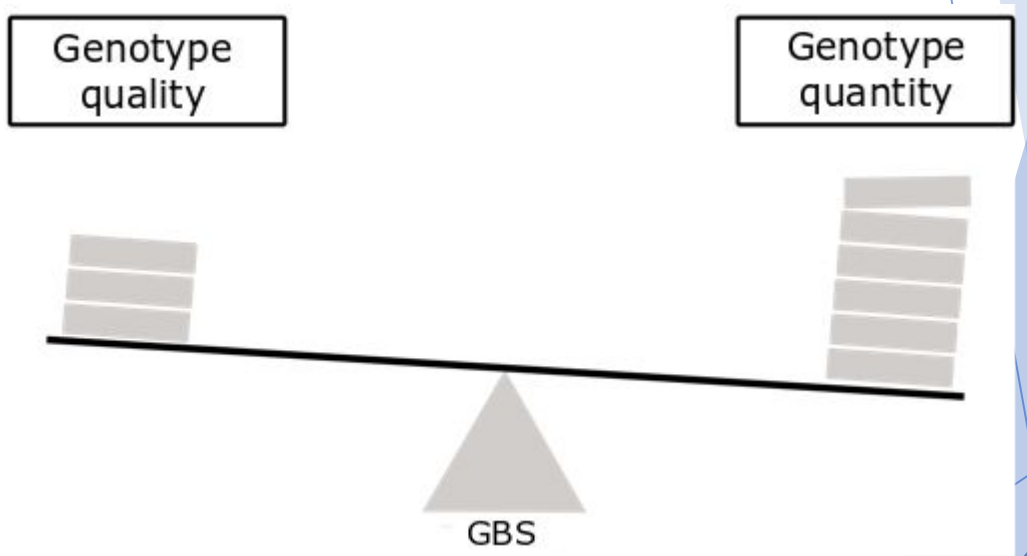
- Short sequences (SNPs, indels)
- Structural variants (number of copies, inversions, translocations)

Molecular markers

- RFLP, RAPD, AFLP, and SSR
- Arrays (For Roses: $$$$$)
- Sequencing (For Roses: $)

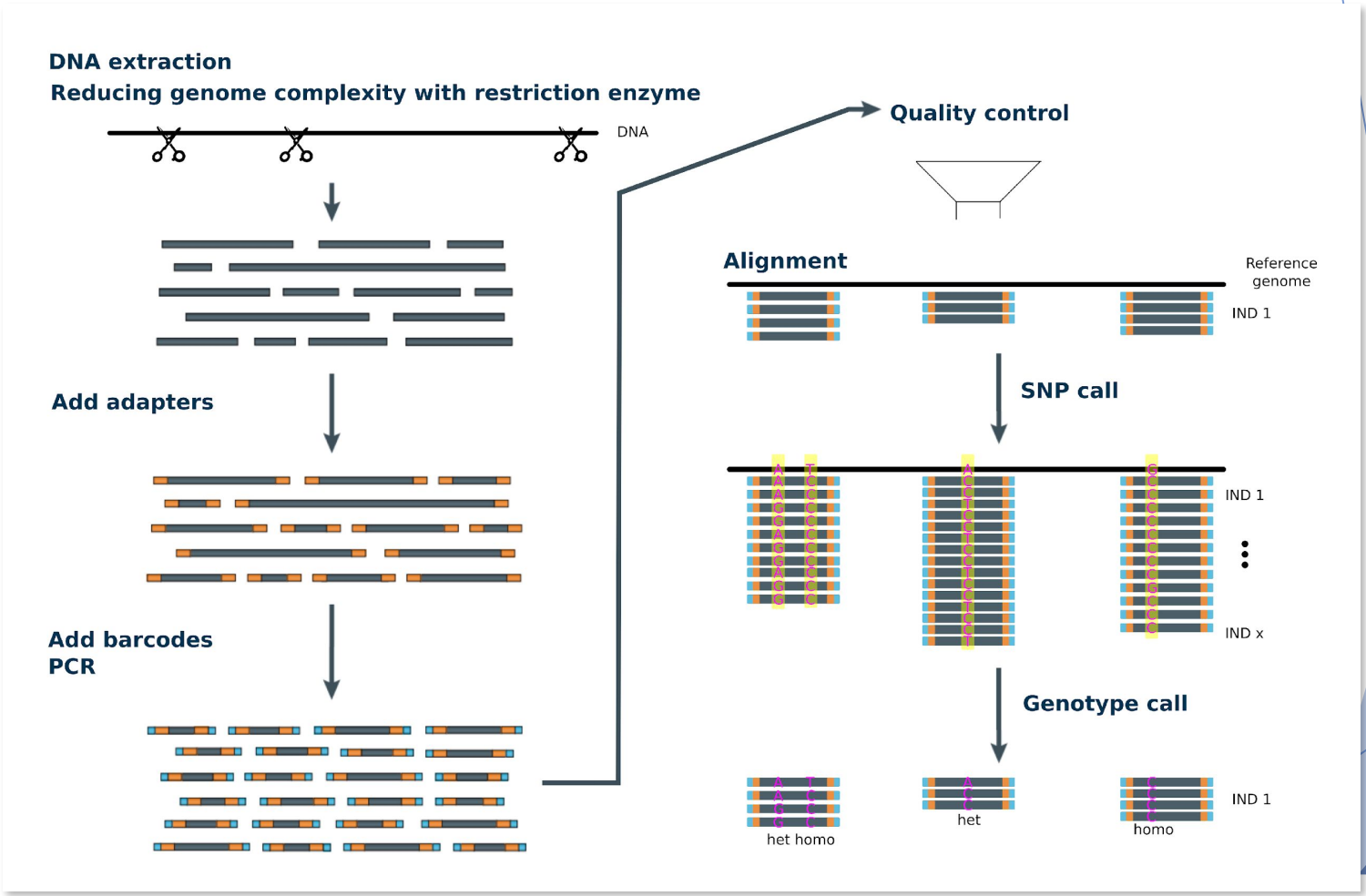
Sequencing Experiment Design

- Study goal
- Sequencer capacity
- Number of individuals per lane
- Number of sequenced loci



TOOLS FOR POLYPLOIDS

GBS Overview

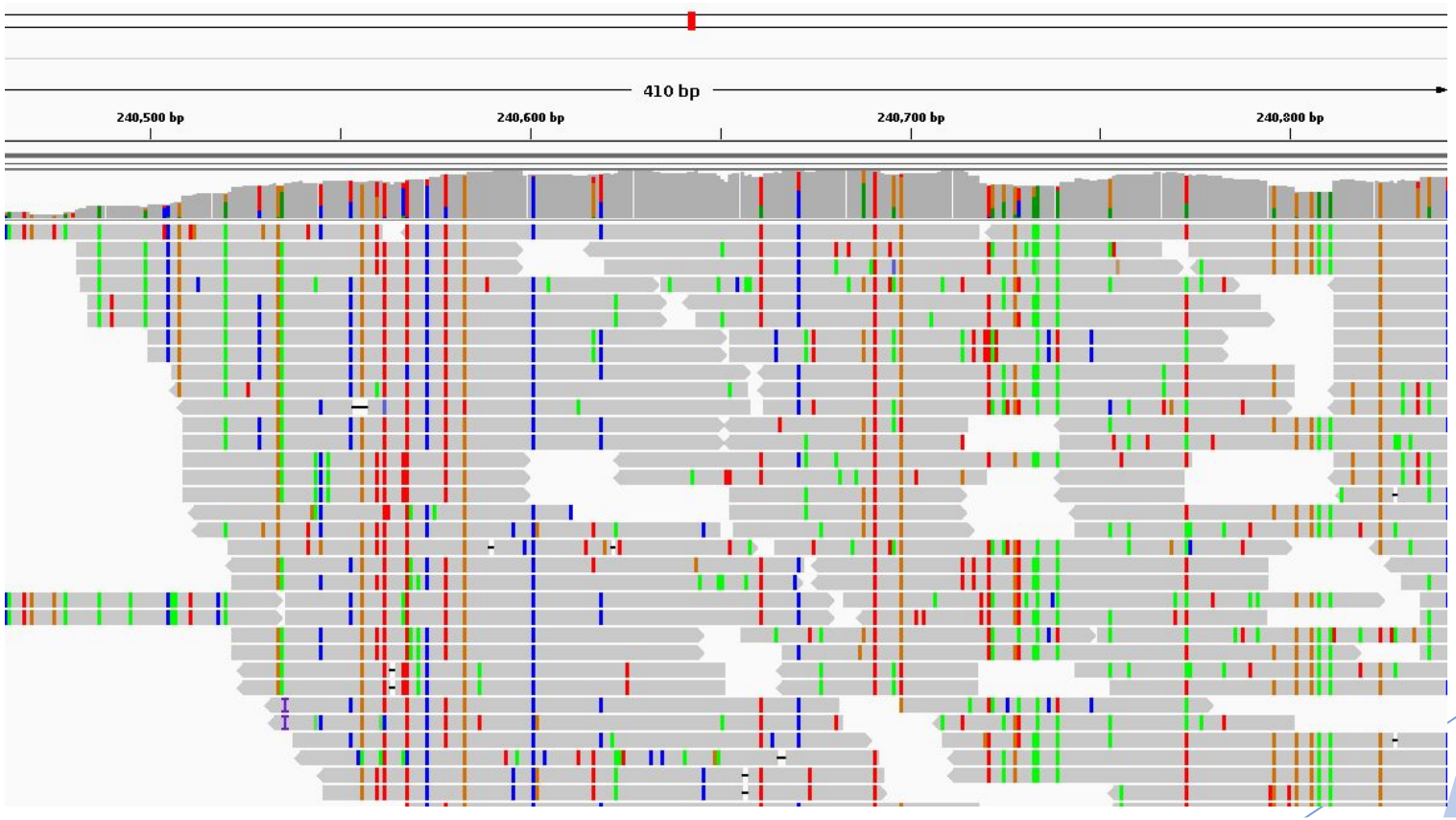


TOOLS FOR POLYPLOIDS

SNP Calling

- Whole Genome Sequencing (WGS)

Image: IGV

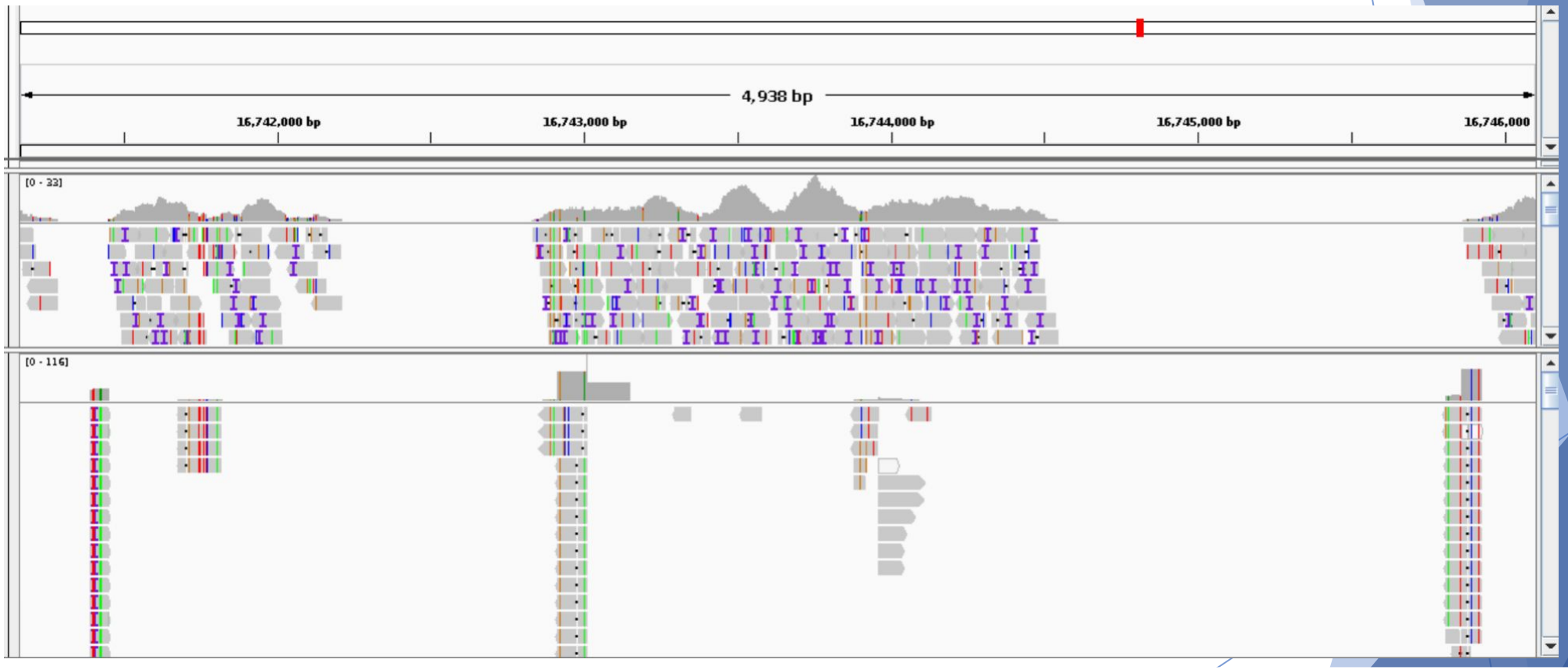


TOOLS FOR POLYPLOIDS

SNP Calling

- Exome sequencing (top) and Genotyping-by-Sequencing (bottom)

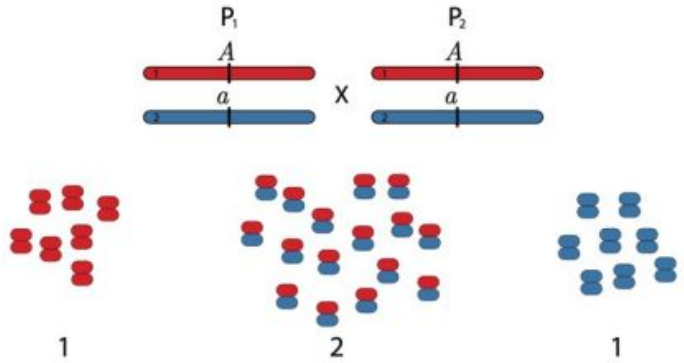
Image: IGV



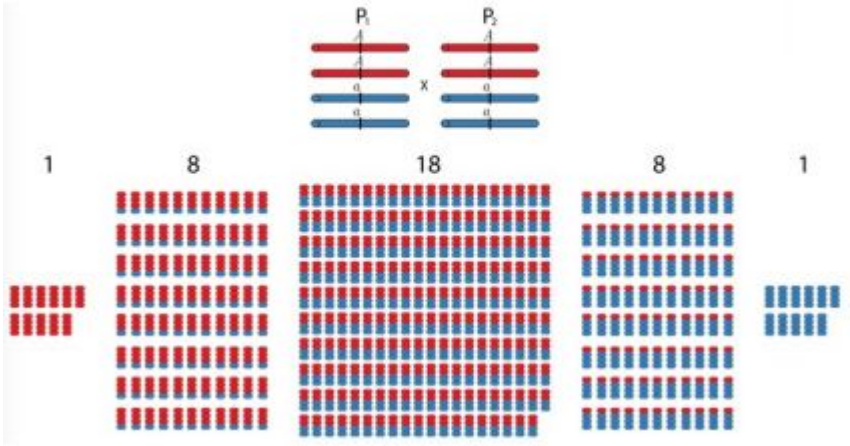
TOOLS FOR POLYPLOIDS

Dosage calling

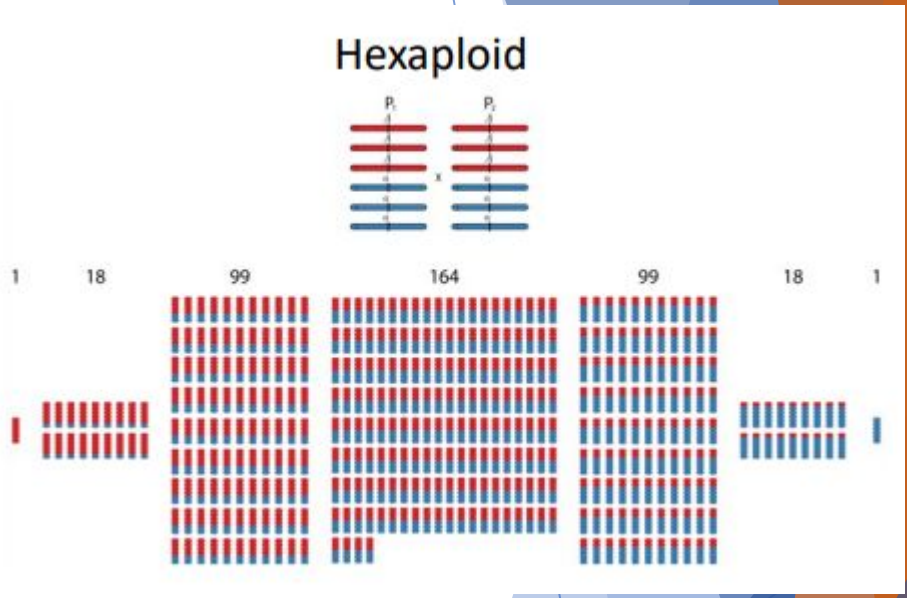
Diploid



Tetraploid



Hexaploid

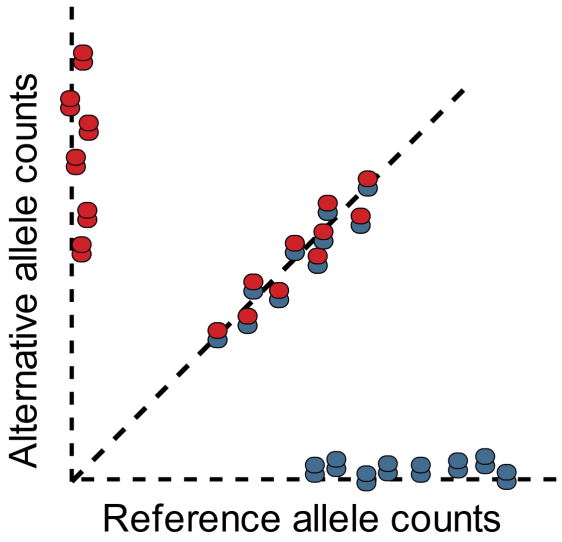


TOOLS FOR POLYPLOIDS

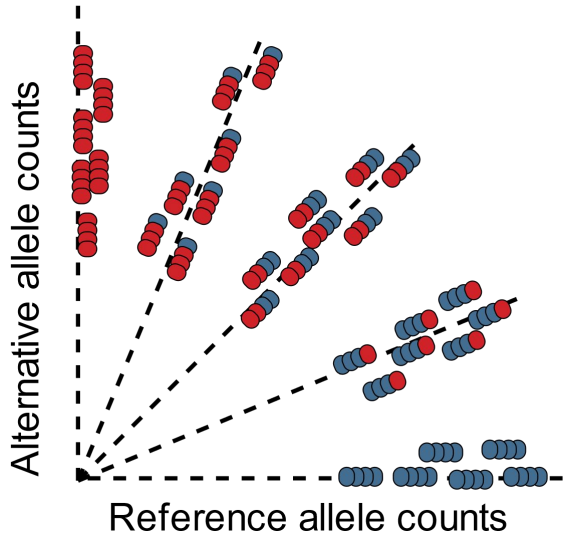
Dosage Calling

- The theory

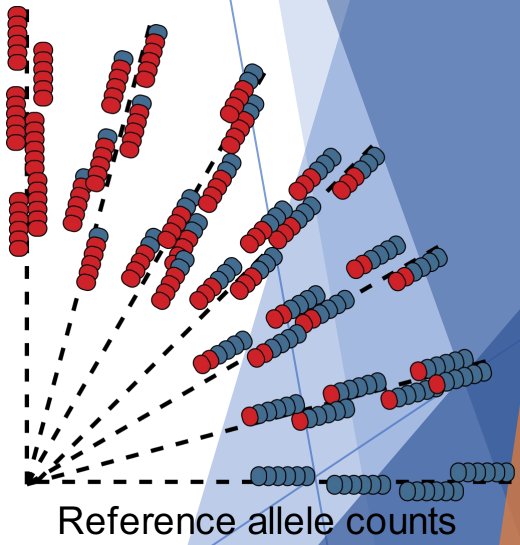
Diploid



Tetraploid



Hexaploid

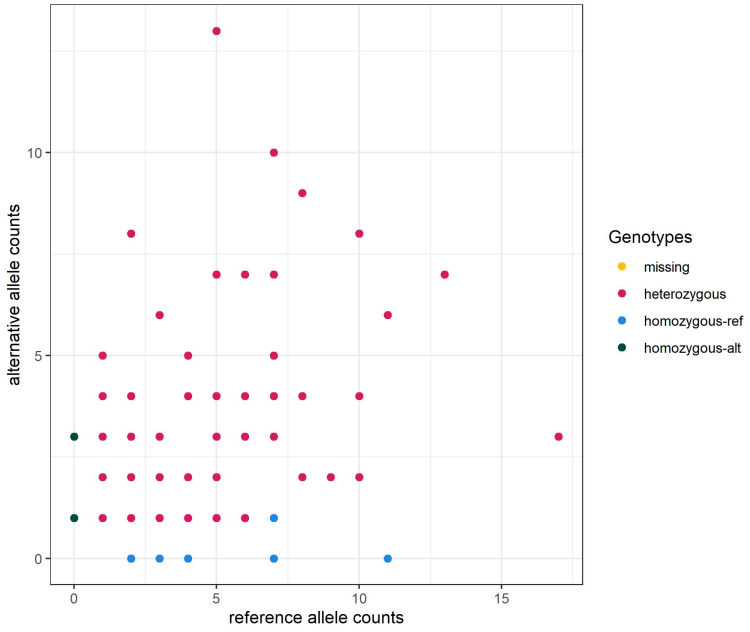


TOOLS FOR POLYPLOIDS

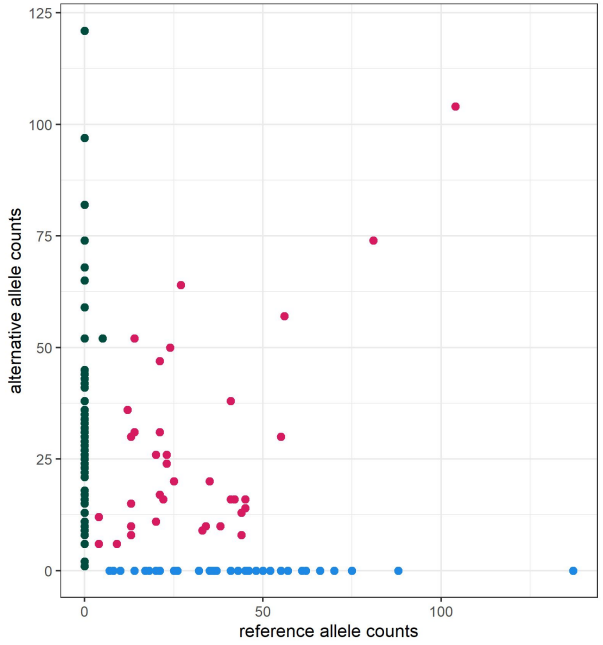
Dosage Calling

- The reality

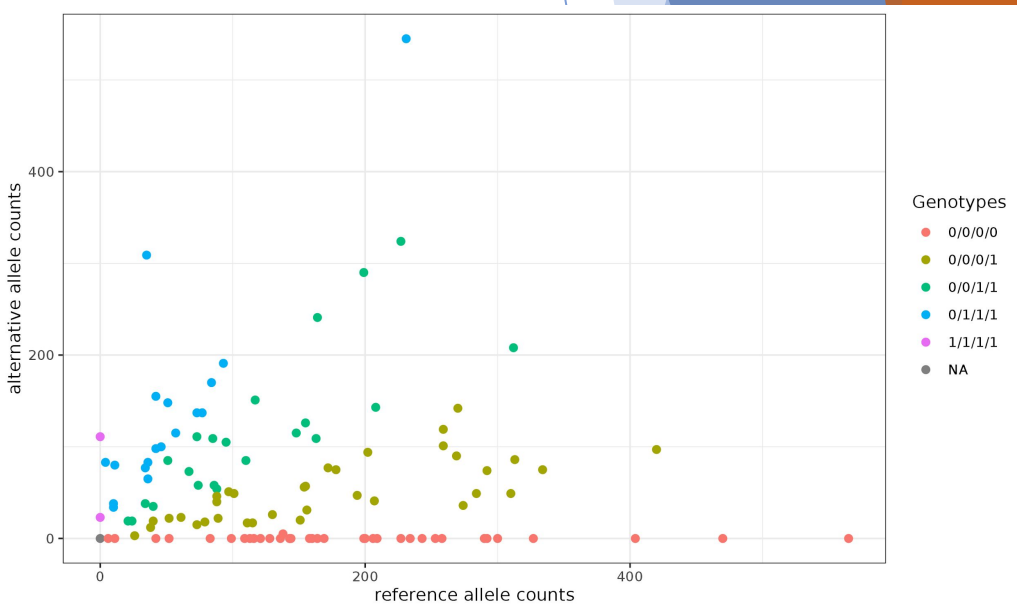
Diploid (mean depth 6)
N = 200
Aa x Aa



Diploid (mean depth 96)
N = 138
Aa x Aa



Tetraploid (mean depth 83)
N = 114
AAaa x AAaa

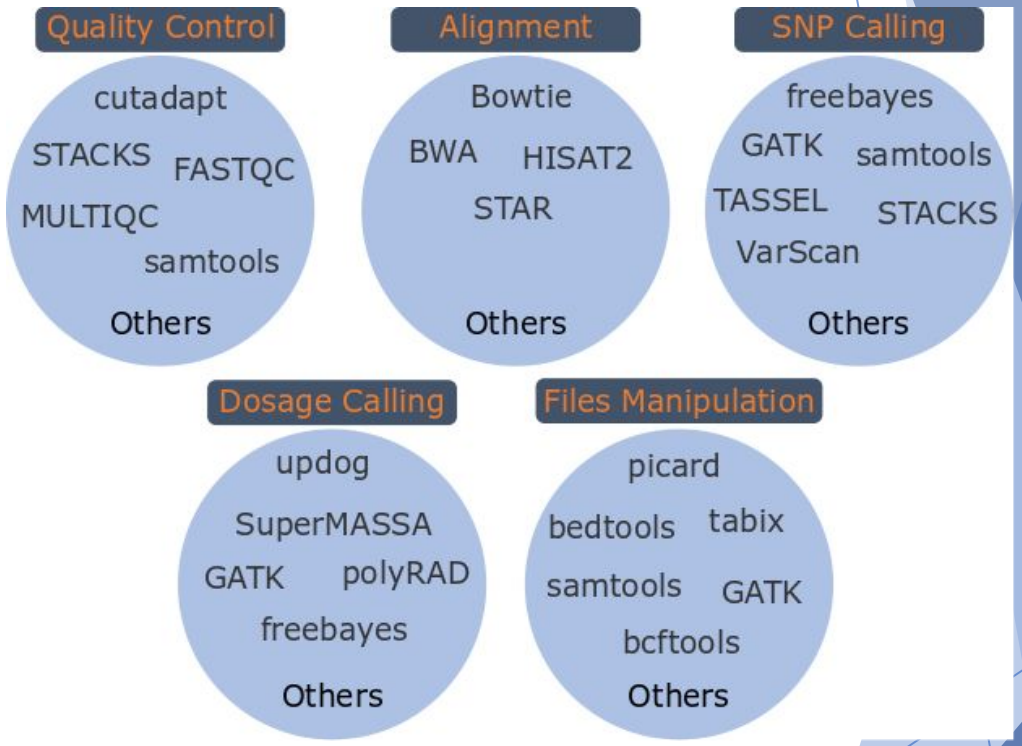


TOOLS FOR POLYPLOIDS

Sequencing Data - Technical Difficulties

- Large files

- Many software
- Many programming languages
- Different Operational Systems
- Updates

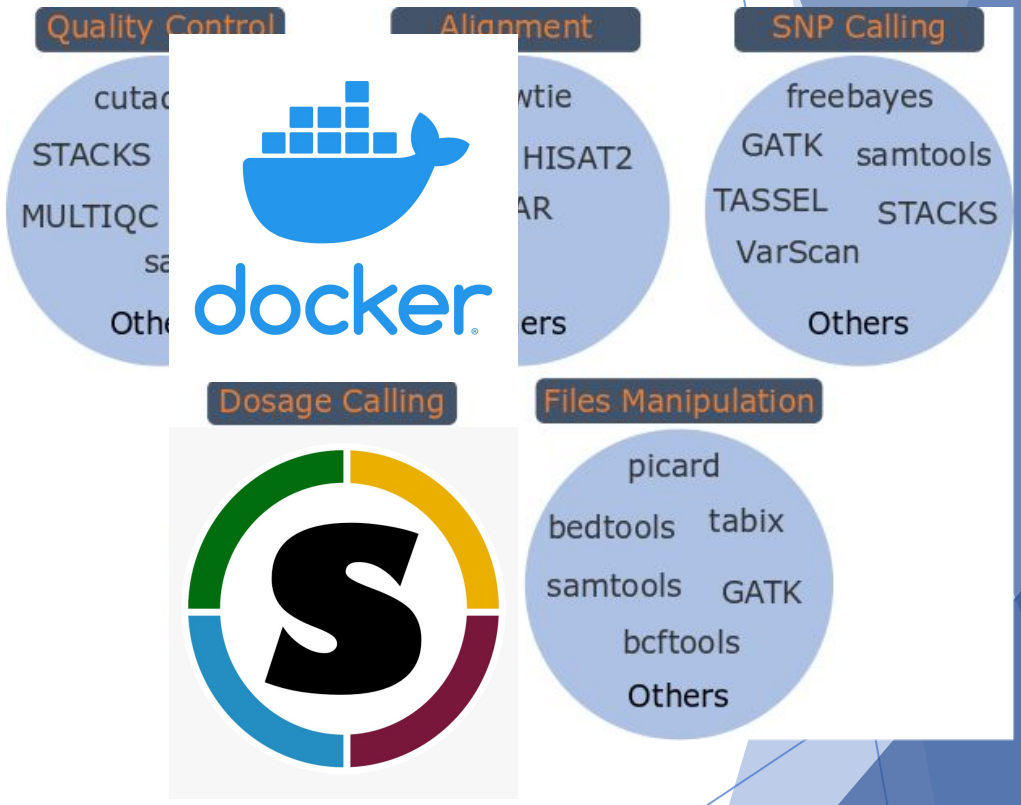


TOOLS FOR POLYPLOIDS

Sequencing Data - Technical Difficulties

- Large files
  - High Performance Computing (HPC)
  - Management systems (SLURM, SGE)
  - Cloud (Google, Amazon)

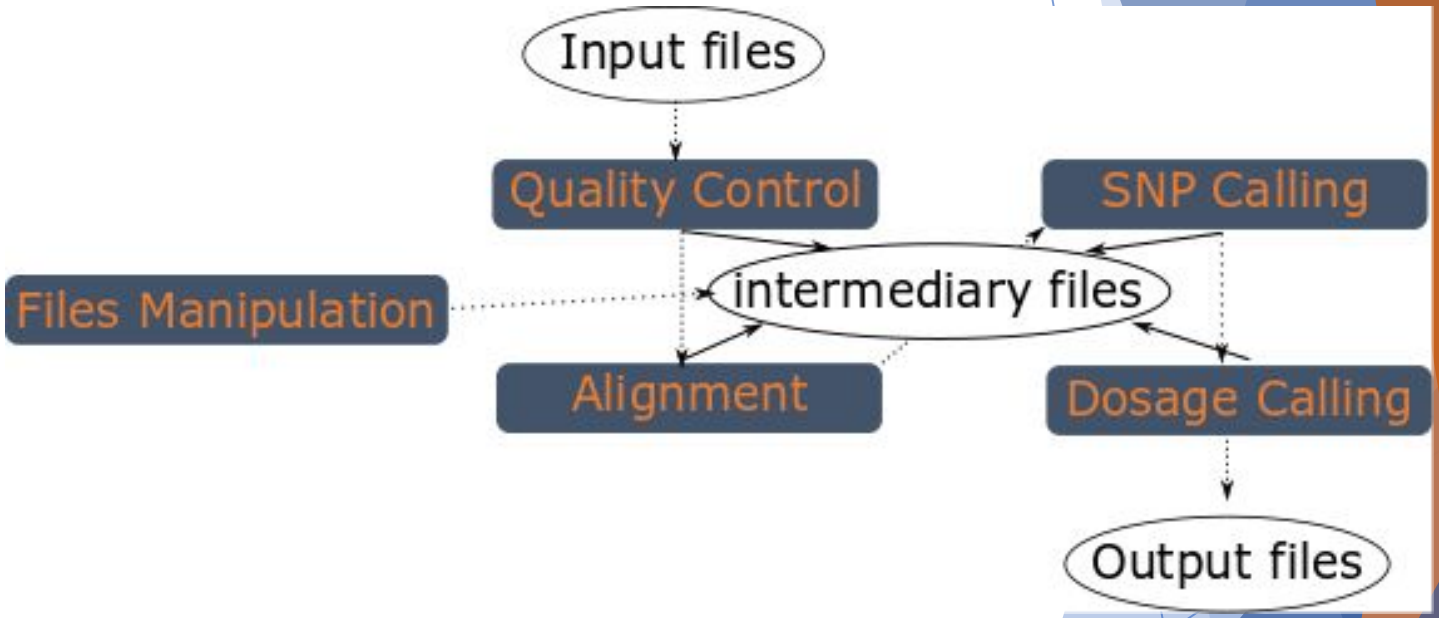
- Many software
- Many programming languages
- Different Operational Systems
- Updates
  - Containers
    - Docker
    - Singularity (usually available in HPC)
    - BioContainers



TOOLS FOR POLYPLOIDS

Sequencing Data - Technical Difficulties

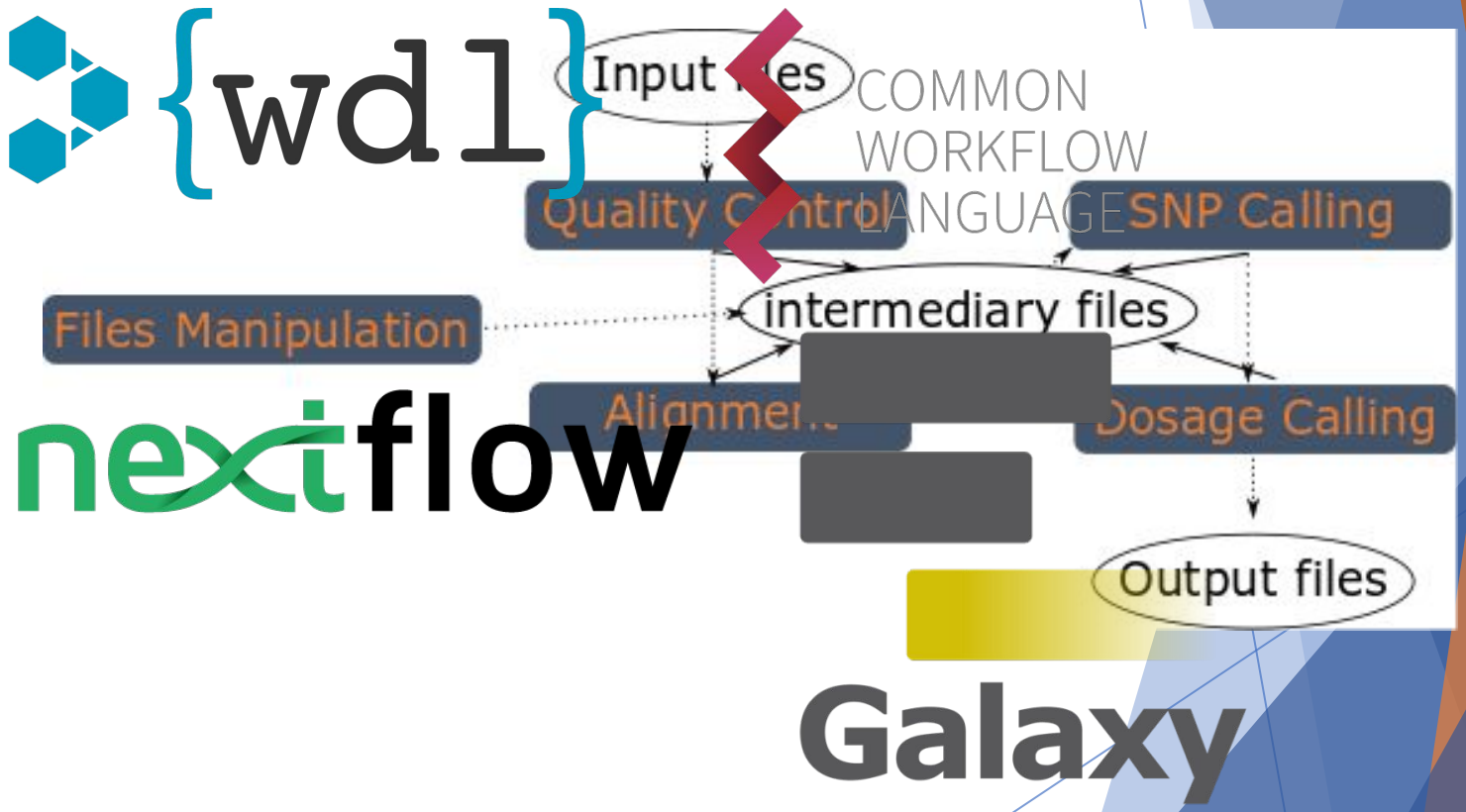
- Many steps
- Many file formats



TOOLS FOR POLYPLOIDS

Sequencing Data - Technical Difficulties

- Many steps
- Many file formats
  - Workflows systems
    - Galaxy
    - Nextflow
    - Snakemake
    - CWL
    - WDL
  - Workflows repositories
    - Dockerstore
    - WorkflowHub
  - Run workflows on Cloud
    - Galaxy
    - DNAnexus
    - Terra
    - AnVIL
    - SevenBridges



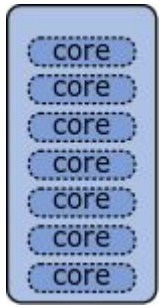
TOOLS FOR POLYPLOIDS

Sequencing Data - Technical Difficulties

- Resources optimization
  - Time
  - Cores
  - Nodes
  - RAM memory

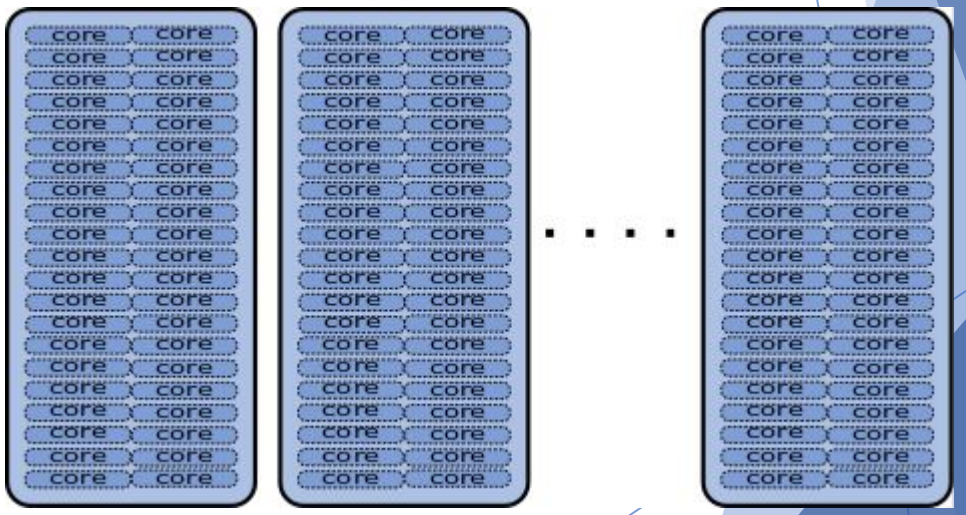
Personal Computer:

4GB RAM; 8 cores; 1 node



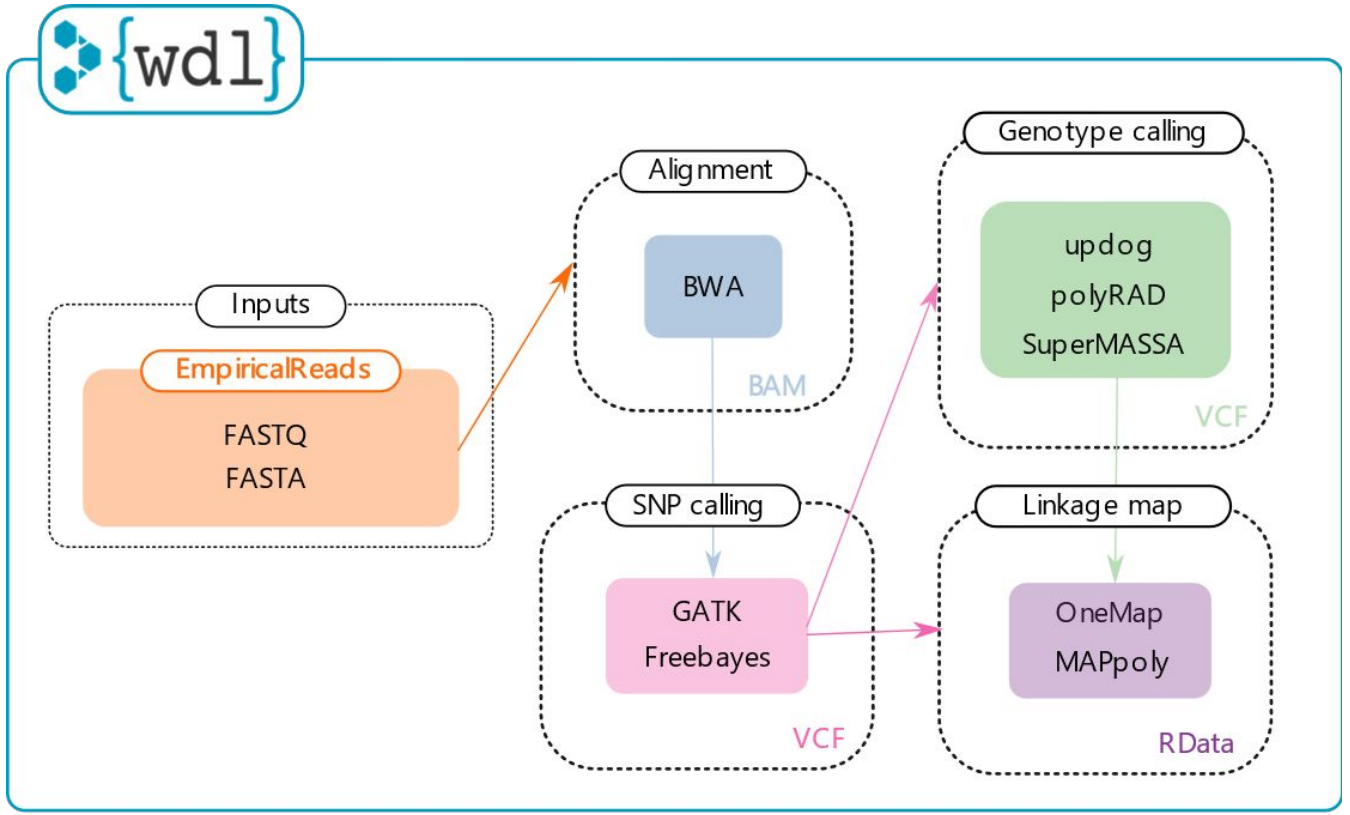
High Performance Computing (Texas A&M):

384GB; 48 cores per node; 900 nodes



TOOLS FOR POLYPLOIDS

Reads2Map

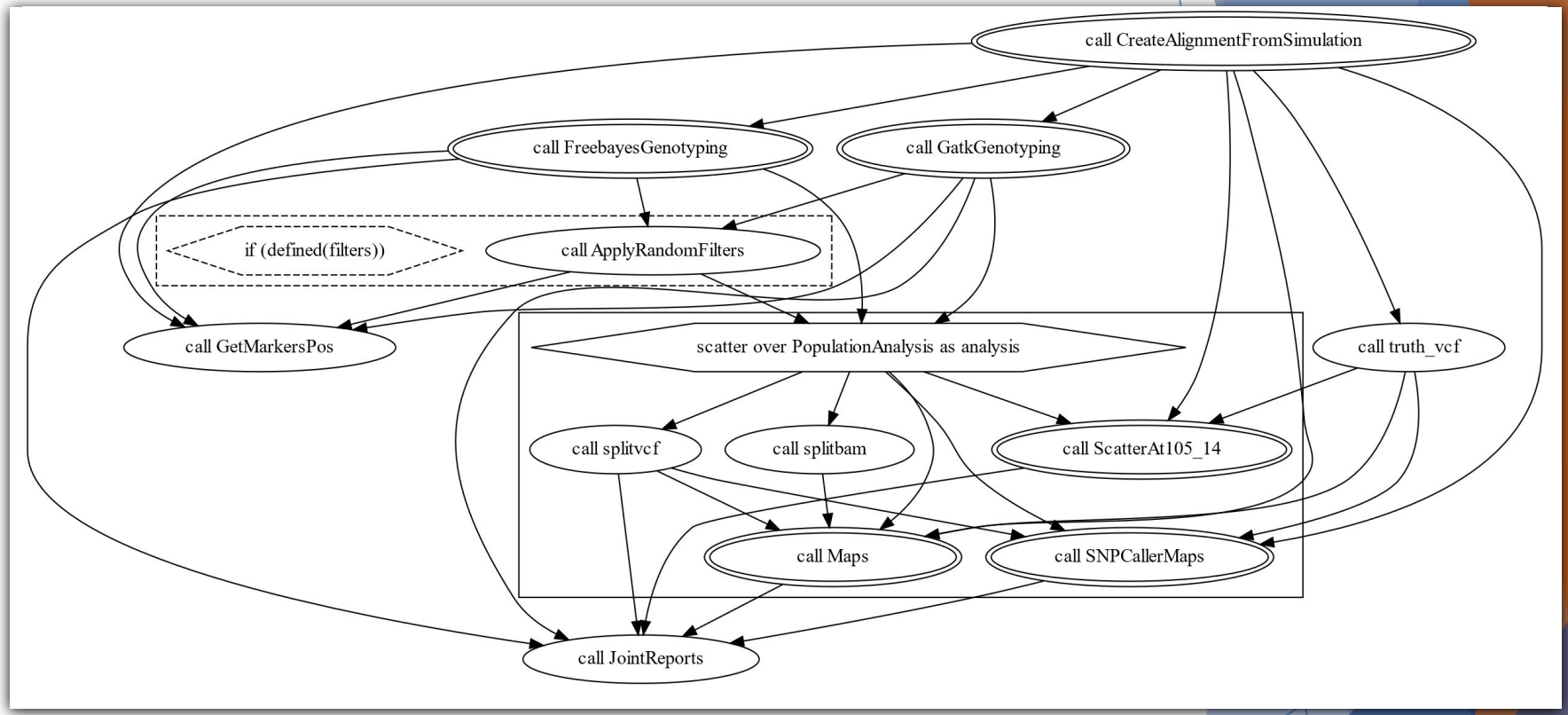


Available in Github, Dockerstore and WorkflowHub

TOOLS FOR POLYPLOIDS

Implementation

- Workflows
  - Sub-workflows
    - Tasks




```
$ java -jar /path/to/womtool.jar graph tasks/SimulatedSingleFamily.wdl > SimulatedSingleFamily.dot
$ dot –Tsvg SimulatedSingleFamily.dot –o SimulatedSingleFamily.svg
```

TOOLS FOR POLYPLOIDS

Implementation

- Cloud environments
  - terra.bio
- High Performance Computing (HPC)
  - Cromwell
  - MiniWDL
  - dxWDL

```
$ java -jar /path/to/cromwell.jar run -i inputs/EmpiricalSNPCalling.inputs.json
EmpiricalSNPCalling.wdl
```

Tutorials

- polyRAD tutorial
- updog tutorial
- fitPoly tutorial
- (TASSEL) Variant and Genotype Calling in Highly Duplicated Genomes (Lindsay Clark)
- Step-by-step of SNP and dosage calling using containers and WDL workflows

TOOLS FOR POLYPLOIDS

References

- Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A.; Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. Molecular Ecology, 22(11), 3124–3140. https://doi.org/10.1111/mec.12354
- Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q.; Buckler, E. S. (2014). TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS ONE, 9(2), 1–11. https://doi.org/10.1371/journal.pone.0090346
- Garrison, E.; Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. ArXiv E-Prints, 9. https://doi.org/1207.3907
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M.; DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research, 20(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

TOOLS FOR POLYPLOIDS

This project is funded by USDA NIFA Specialty Crop Research Initiative Award
2020-51181-32156 (09/01/20 - 08/31/24)

References

- Gerard, D., Ferrão, L. F. V., Garcia, A. A. F.,& Stephens, M. (2018). Genotyping Polyploids from Messy Sequencing Data. Genetics, 210(3), 789-807. doi: 10.1534/genetics.118.301468.
- Wadl, P. A., Olukolu, B. A., Branham, S. E., Jarret, R. L., Yencho, G. C.; Jackson, D. M. (2018). Genetic Diversity and Population Structure of the USDA Sweetpotato (Ipomoea batatas) Germplasm Collections Using GBSpoly. Frontiers in Plant Science, 9, 1166. https://doi.org/10.3389/fpls.2018.01166
- Serang, O., Mollinari, M.; Garcia, A. A. F. (2012). Efficient exact maximum a posteriori computation for bayesian SNP genotyping in polyploids. PLoS ONE, 7(2), 1–13. https://doi.org/10.1371/journal.pone.0030906
- Clark, L. v., Lipka, A. E.; Sacks, E. J. (2019). polyRAD: Genotype Calling with Uncertainty from Sequencing Data in Polyploids and Diploids. G3: Genes|Genomes|Genetics, 9(March), g3.200913.2018. https://doi.org/10.1534/g3.118.200913

TOOLS FOR POLYPLOIDS

This project is funded by USDA NIFA Specialty Crop Research Initiative Award
2020-51181-32156 (09/01/20 - 08/31/24)

Project Members

TEXAS A&M UNIVERSITY

Plant & Food Research
Rangahau Ahumāra Kai

Cornell University

WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON

PennState

WAGENINGEN
UNIVERSITY & RESEARCH

WASHINGTON STATE UNIVERSITY

UNIVERSITY OF MINNESOTA
Driven to Discover®

NC STATE UNIVERSITY

UNIVERSITY OF ARKANSAS

THE UNIVERSITY OF MAINE

Oregon State University

TOOLS FOR POLYPLOIDS

This project is funded by USDA NIFA Specialty Crop Research Initiative Award # 2020-51181-32156 (09/01/20 - 08/31/24)

Other Collaborators

Genetwister

DAICEL arbor biosciences

VacCAP

Neuhouse Farms

bejo

Spring Meadow

Royal Van Zanten

American Garden Rose Selections™

Plant Sciences PSI Inc. Advancing Agriculture through Science

IITA Transforming African Agriculture

CIAT

CIP INTERNATIONAL POTATO CENTER

Potatoes USA

Woolf Roses L.L.C.

star® ROSES and PLANTS

Dekker chrysanten

BAILEY®

CSS FARMS

Weeks Roses

DÜMMEN ORANGE

illumina®

PACIFIC BERRY BREEDING

DELIFLOR Chrysanten

genVATIONS

Driscoll's

TOOLS FOR POLYPLOIDS

This project is funded by USDA NIFA Specialty Crop Research Initiative Award # 2020-51181-32156 (09/01/20 - 08/31/24)